26.4 A 0.13µm 2.125MB 23.5ns Embedded Flash with 2GB/s Read Throughput for Automotive Microcontrollers

Christoph Deml, Maciej Jankowski, Carmen Thalmaier

Infineon Technologies, Neubiberg, Germany

Automotive real-time applications such as powertrain and motor management rely on non-volatile storage of increasing amounts of program code and data. They require memories with a high bandwidth and low latency. External memories suffer from speed limitations due to I/O circuits, pads and wiring on PCBs. Therefore high-end 32 bit microcontrollers use on-chip flash for maximum data throughput.

In this paper, an embedded flash memory fabricated in 0.13µm CMOS with a nonvolatile technology extension is presented. The module is specified for an operating junction temperature range of -40 to 150°C using a twin supply of 1.5V and 3.3V each ±5%. Reliability requirements are fulfilled by employing single-bit correction plus double-bit detection (72, 64, 4) ECC logic residing in the host system. A die micrograph is shown in Fig. 26.4.7.

Using a 0.54×0.7µm² uniform channel programming (UCP) cell in a selected NOR architecture, the flash module has an area of 23.4mm². It contains a 2MB program code bank composed of seven 256KB sectors, one 128KB sector and two 64KB sectors, with the latter two erasable in 16KB blocks. Additionally 16KB sector is used for firmware, chip and customer configuration data. A single program memory access provides 288b within four clock cycles, allowing a continuous 72b per cycle burst read. Additionally , there are two 64KB banks for data storage that concurrently deliver 72b each with the same access time. These three memory banks share a state machine and common voltage generation circuits supporting parallel read, erase and program operations.

The operation of a flash read access is split into address information distribution, wordline (WL) charging, redundancy activation decision, source line (SL) discharging, bitline (BL) settling and sensing, and finally propagation on data output lines. For maximum performance each part has to be optimized.

Each sense amplifier corresponds to a slice that is 16 local BLs wide and 11 sectors high. Read performance is optimized by minimization of capacitive load at the sense amplifier input. Therefore only one at a time out of 176 local BLs is connected to the sense amplifier via a global BL. The fastest address-dependent connection of a local-to-global BL is achieved without level shifting. The use of thin-gate-oxide transistors necessitates protecting them against high voltages by thick-gate-oxide transistors in series during erase and program.

Shrinking the feature size causes only a slight variation in the capacitance per length of interconnects and scales the resistance per length with its reciprocal square. To minimize delay, long address and data lines are routed in higher metal layers, which provide a larger cross sectional area. The global BL is separated in top metal with an empty layer beneath, as shown in Fig. 26.4.1. The relatively high resistive poly WL is stitched to a parallel metal WL at eight points in each memory wing to reduce resistance.

The cumulative resistance of all transistors between the sense amplifier and cell also has a direct influence on the BL settling time. Minimizing the BL setting time by increasing transistor width incurs an area penalty and is therefore a subject for trade off.

Current sensing is chosen for the read-out because it is more robust to mismatch and noise immunity than latch-type voltage sensing [1] and does not require timing of sensing phases, which may prove critical [2]. During read, the BL voltage must be regulated while the SL is connected to ground. In conventional schemes, BL and SL reside at ground potential while unselected, which necessitates a wide BL voltage swing at the start of selection with a well-controlled end voltage. In conjunction with propagation on the BL this is a slow process. To overcome this deficiency, a continuous precharge concept was developed [3]. The reading process is shown in Fig. 26.4.2. When unselected, local BLs as well as local SLs reside precharged at the operating point of the sense amplifier using a separate precharge buffer. Upon selection, the precharge buffer is disconnected from the local SL and local BL. The local SL is simply connected to ground and the local BL is connected to the sense amplifier. Capacitive coupling from the discharging local SL to the local BL causes a small voltage difference at the BL, which is quickly equalized by the sense amplifier.

The cell current is influenced by die-to-die process deviations, voltage supply variations and temperature. The reference current must track this behavior to guarantee robustness and speed optimization. Tracking is achieved by the use of cells in a separate array as a reference. The reference-generation circuit of Fig. 26.4.3 is built strictly symmetrical to the BL decoder and sector. The trip point is adjusted by selection of n at the n:8 reference current mirror. Parasitic capacitances at node N1 transform the difference of cell and reference current to a voltage swing. Therefore the sensing delay equals the integration time corresponding to a voltage swing from one rail to the inverter trip point. This delay is minimized by transistor P1, which limits the voltage at node N1 to approximately $2V_t$.

The WL decoder of Fig. 26.4.4 includes a level shifter to transform the 1.5V selection signal to the regulated 3.3V WL selection level. A thick-gate-oxide pass-transistor protects the driver from the high voltage during erase and program.

The circuit architecture shown in Fig. 26.4.5 combines the described solutions. Multiplexing 16 local BLs to one global BL during read is realized by the global-to-local switch. Said separation of global BLs is only possible if the number of global SLs for programming is reduced. This is achieved by multiplexing two local to one global SL thus implementing two 288B pages per WL.

Redundancy activation is controlled by an asynchronous content-addressable memory (CAM) that compares the address with known error location addresses. The CAM is a full custom design optimized for area and speed. The result of the comparison is processed by a distributed redundancy decoder, which is integrated into the BL decoder.

The access time in Fig. 26.4.6 is mainly dependent on the address transition (e.g., local BL or WL hit, WL or sector jump), sector size, redundancy usage, data transition, supply voltages, temperature and device parameters. A measurement at worst case supply voltages and 150°C ambient temperature confirms full system operation at a clock frequency of 170MHz, which corresponds to a random access time of less than 23.5ns.

The clock frequency validated in this work results in a 1.3× higher data rate compared to previous work [4], which also uses a four cycle burst transfer with 72b each. Moreover, our flash is able to read all three memory banks in parallel. Because of the continuous precharge concept, the module achieves 2GB/s burst read throughput.

Acknowledgements:
We thank Dr. Anja Dübotzky, Gaby Haack and Ute Rossberg for preparation of the micrographs.

References:
[1] B. Wicht, *Current Sense Amplifiers for Embedded SRAM in High-Performance System-on-a-Chip Designs,* Springer, pp. 16-37, 2003.
[2] D. Farenc et al., "12ns Random Access Time in Pipeline Mode for High Performance Embedded FLASH Applications in a 0.13µm Technology," *IEEE Nonvolatile Semiconductor Memory Workshop,* pp. 67-68, Aug., 2004.
[3] C. Deml et al., "Precharge Arrangement for Read Access for Integrated Nonvolatile Memories," *United States Patent and Trademark Office,* US 2005/0128813 A1, Jun. 16, 2005.
[4] Freescale Semiconductor, "MPC5554 and MPC5553 Microcontroller Reference Manual," Rev. 3.1, Nov. 9, 2005.

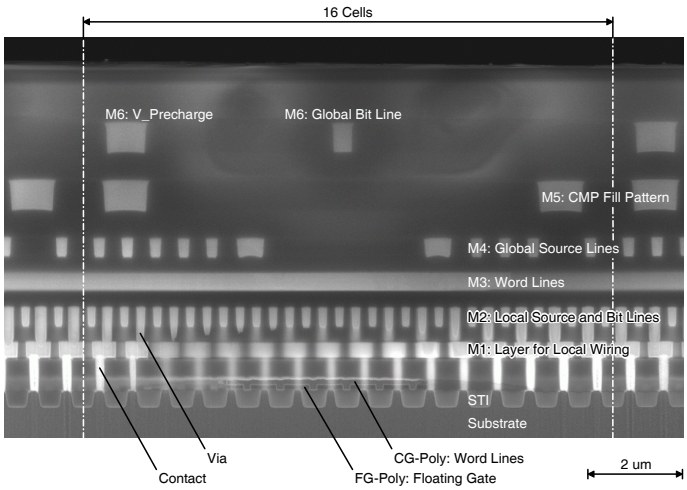


Figure 26.4.1: Cross-section of cell array with an angle of 3˚ relative to the WL to show the WL and contacts in one micrograph.

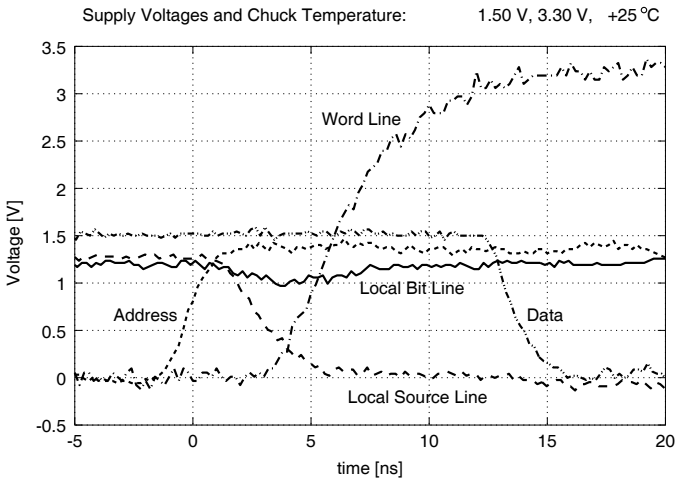


Figure 26.4.2: Measured single read access at nominal conditions.

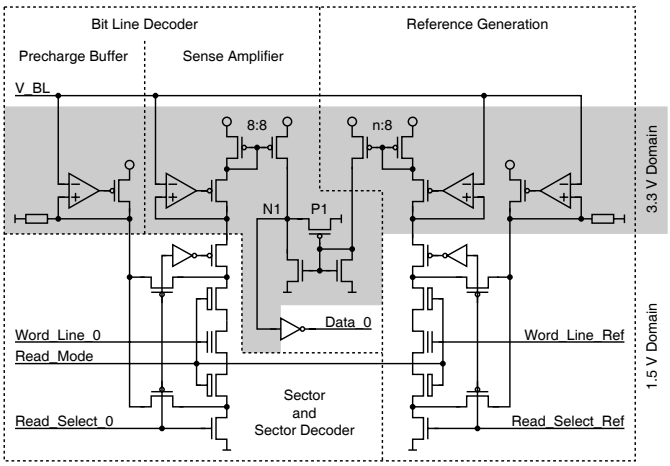
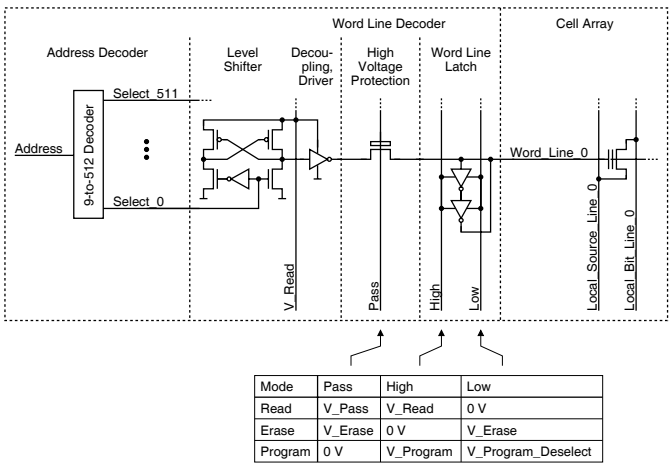


Figure 26.4.3: Bitline decoder and reference generation.



Mode	Pass	High	Low
Read	V_Pass	V_Read	0 V
Erase	V_Erase	0 V	V_Erase
Program	0 V	V_Program	V_Program_Deselect

Figure 26.4.4: Word line decoder.

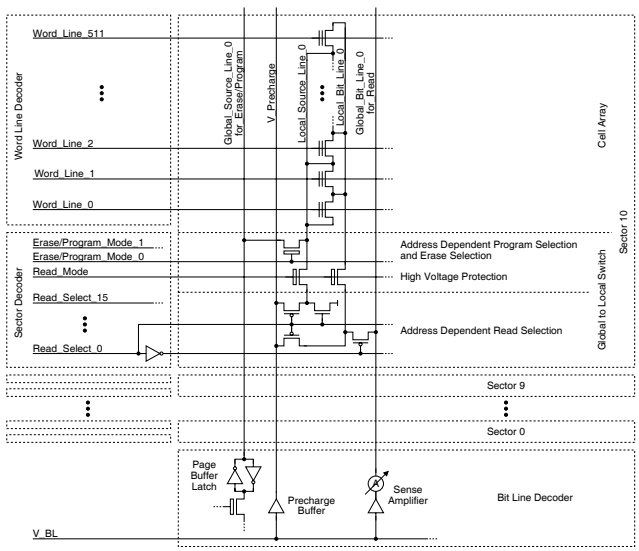


Figure 26.4.5: Program memory architecture and global to local switch.

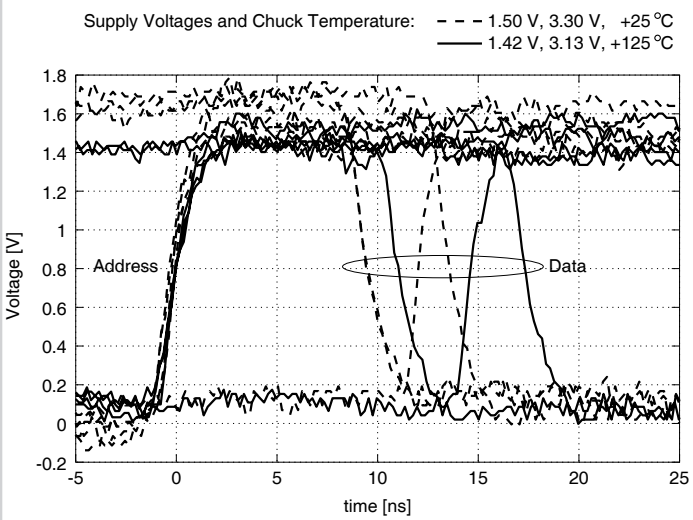


Figure 26.4.6: Measured read accesses including the cases for minimum and maximum access time.

Continued on Page 617

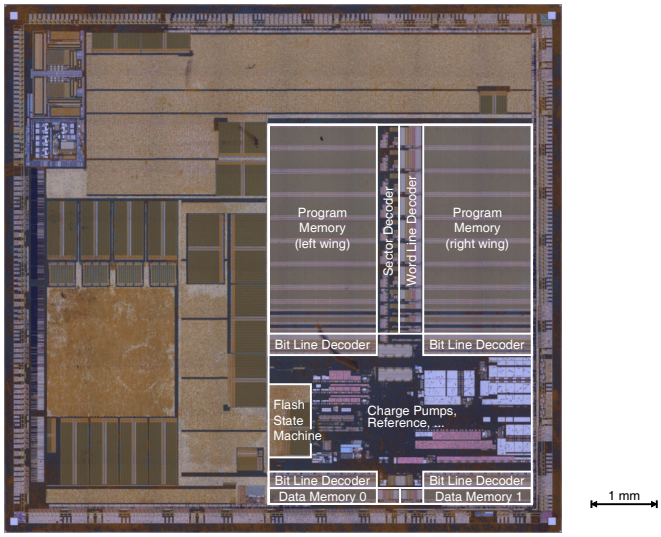


Figure 26.4.7: Chip micrograph (metallization removed).

